

---

---

# Towards Realistic Transfer Learning Methods: Theory and Algorithms

---

---

*A thesis submitted in fulfilment of the requirements  
for the degree of*

Doctor of Philosophy  
*in*  
Computer Science

*by*

**Feng Liu**

*to*

School of Computer Science  
Faculty of Engineering and Information Technology  
Australian Artificial Intelligence Institute  
University of Technology Sydney  
NSW - 2007, Australia  
2020



## CERTIFICATE OF ORIGINAL AUTHORSHIP

**I**, *Feng Liu*, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the School of Computer Science at the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE:

Production Note:  
Signature removed prior to publication.

DATE: 08, Sep., 2020

## ABSTRACT

**T**ransfer learning aims to leverage knowledge from domains with abundant labels (i.e., source domains) to help train a classifier or predictor for the domain with insufficient labels (i.e., target domain). The trained classifier or predictor is expected to have better performance (e.g., higher accuracy) than classifiers only trained with data in the target domain.

Although recent research of transfer learning has shown a decent ability to transfer knowledge from a source domain to a target domain, most research require certain assumptions to ensure their efficacy. These assumptions are probably not realistic, which means that existing transfer learning methods still face several unsolved and challenging problems in real world.

This thesis aims to address four orthogonal problems faced by existing transfer learning methods: 1) How to test if feature spaces of two domains are from different distributions; 2) How to transfer knowledge when labels in the source domain cannot be perfectly annotated (i.e., the source domain contains noisy labels); 3) How to transfer knowledge when source and target domains have different dimensions (i.e., heterogeneous scenario); and 4) How to transfer knowledge across multiple source domains and a different-dimension target domain.

To address Problem 1), this thesis presents two new two-sample tests to test

---

if the feature spaces of source domains and target domain are from different distributions. One is suitable for low-dimension data (Chapter 3) and another for high-dimension data (Chapter 4). If feature spaces of domains are statistically different, we need to use transfer learning methods on these domains. Moreover, the test statistics used in the proposed tests can be used to measure the distributional discrepancy between two domains.

To address Problem 2), this thesis presents a theoretical bound to show that existing transfer learning methods cannot work well when a source domain contain noisy labels. Then, a novel transfer learning approach is proposed to transfer knowledge across a source domain (with noisy labels) and a target domain. Finally, a generalization bound is proved to explain why the proposed method can reliably transfer knowledge across domains in noisy scenario (Chapter 5).

To address Problem 3), the most challenging problem in the field of domain adaptation, Chapter 6 presents a theorem to show when we can reliably transfer knowledge across two different-dimension (i.e., heterogeneous) domains and propose a solution to this problem. Since methods in Chapter 6 assume that the number of samples in two domains must be the same (i.e., two balanced domains), Chapter 7 presents a novel fuzzy-relation based method to transfer knowledge across two imbalanced domains.

To address Problem 4), Chapter 8 presents a novel fuzzy-relation neural network to transfer knowledge from multiple source domains to a target domain, where any of two domains are heterogeneous (i.e., feature spaces of any of two domains have different dimensions).

To conclude, this thesis not only propose a set of effective methods for realistic transfer learning, but also contribute to theory of transfer learning.

## DEDICATION

*To my loving wife, parents and families...*

## ACKNOWLEDGMENTS

It is a memorial and exciting journey at University of Technology Sydney (UTS) for pursuing my Ph.D. degree in the past three and half years. I am sincerely grateful to the people who inspired and helped me in many ways.

I would like to express my foremost and deepest gratitude to my principal supervisor, A./Professor Guangquan Zhang. Without his patience and encouragement, I would not have been able to complete this Ph.D. program. He taught me step by step how to become a qualified researcher from its beginning. He always led me to the right research direction with his expert knowledge of theory and abundant research experience. He placed considerable trust in my research ability and unconditionally support me in pursuing my own research interests. His wisdom and immense knowledge always enlightened me to go further and deeper in my research. Without his critical comments, I would waste my time on trivial research ideas. Discussion with him greatly improves the scientific aspect and quality of my research. He helped me to build my confidence in my research outcomes and to be hopeful when faced with any difficulty, from academic to living.

Meanwhile, I am greatly indebted to my co-advisor, Distinguished Professor

---

Jie Lu. Her decisiveness and sharp insights continuously motivated me when I got lost or afraid about the future. Her confidence and enthusiasm inspired me to do the right thing even when the road got tough. I felt extremely honored to be guided by such a rigorous researcher as well as an enthusiastic mentor. What she taught me and what I learned from her in the past four years has benefited my Ph.D. study and will be a great treasure throughout my life.

During my Ph.D. period, I am very fortunate to join the Imperfect Information Learning Team as a research intern at RIKEN Center for Advanced Intelligence Project (RIKEN-AIP), working with Prof. Masashi Sugiyama, Dr. Gang Niu and Dr. Bo Han; to join the Gatsby Computational Neuroscience Unit at UCL as a visiting scholar, working with Prof. Arthur Gretton, Dr. Dougal J. Sutherland and Wenkai Xu. I would like to express my thankfulness to Prof. Masashi Sugiyama and Prof. Arthur Gretton that they led me to the machine learning field. Discussion and cooperation with them helped me to deeply understand what the top-tier machine-learning work should be and what characteristics a machine-learning researcher should have. I am impressed and inspired by their persistence, rigour and passion on research.

I would like to express my thankfulness to every member of the Decision Systems & e-Service Intelligence Lab (DeSI) in the Centre for Artificial Intelligence (CAI). It was a wonderful experience to spend four years with these dedicated researchers. I especially thank Dr. Hua Zuo, Dr. Junyun Xuan, Dr. Zheng Yan, Dr. Yi Zhang, Dr. Anjin Liu, Dr. Fujin Zhu, Dr. Feng Gu, Dr. Ning Lu and Zhen Fang who provided insightful comments related to my research problem during my Ph.D. candidature; Dr. Ximeng Wang, Dr. Guanjin Wang, Qian Liu, Dr. Chenlian Hu, Bin Zhang and Bin Wang who have shared their opinions and comments



---

with me, Dr. Anjin Liu, Dr. Junyun Xuan, Hang Yu and Ruiping Yin who shared my joys and sadness.

I must thank Dr. Gang Niu and Dr. Bo Han for the valuable suggestions for my Ph.D. study. Their persistence, rigour and passion on research impressed and inspired me. I have learned much about how to do an excellent machine-learning research from them. I also express my thankfulness to Dr. Dougal J. Sutherland who patiently guides me how to do an excellent machine-learning research. I am also impressed and inspired by his persistence, rigour and passion on research.

Meanwhile, I genuinely thank Jemima Moore, Sue Felix and Michele Mooney for polishing the language of my publications. They are always patient to all my emails of questioning revised sentences. I thank all my wonderful friends, classmates and colleagues for every enjoyable moment.

Last, I would like to express my heartfelt appreciation and gratitude to my wife, parents, and families for their love and support.

## LIST OF PUBLICATIONS

1. **Feng Liu\***, Wenkai Xu\*, Jie Lu, Guangquan Zhang, Arthur Gretton, Dougal Sutherland. Learning Deep Kernels for Nonparametric Two Sample Test. *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, online, 13-18 July, 2020. (\*Equal Contribution). [ERA: A, CORE: A\*]
2. **Feng Liu**, Guangquan Zhang, Jie Lu, Heterogeneous Domain Adaptation: An Unsupervised Approach, *IEEE Transactions on Neural Networks and Learning Systems (IEEE-TNNLS)*, 2020. DOI: 10.1109/TNNLS.2020.2973293. [ERA&CORE: A\*, JCR Q1]
3. **Feng Liu**, Guangquan Zhang, Jie Lu, Multi-source heterogeneous unsupervised domain adaptation via shared-fuzzy-equivalence-relation neural networks, *IEEE Transactions on Fuzzy Systems (IEEE-TFS)*, 2020. DOI: 10.1109/TFUZZ.2020.3018191. [ERA&CORE: A\*, JCR Q1]
4. **Feng Liu**, Guangquan Zhang, Jie Lu, Unsupervised Heterogeneous Domain Adaptation via Shared Fuzzy Equivalence Relations, *IEEE Transactions on Fuzzy Systems (IEEE-TFS)*, Vol. 26, no. 6, pp. 3555-3568, 2018. [ERA&CORE: A\*, JCR Q1]

- 
5. **Feng Liu**, Jie Lu, Bo Han, Gang Niu, Guangquan Zhang, Masashi Sugiyama. Butterfly: One-step Approach towards Wildly Unsupervised Domain Adaptation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (IEEE-TPAMI). [ERA&CORE: A\*, JCR Q1] (Revise&Resubmit)
  6. **Feng Liu**, Jie Lu, Bo Han, Gang Niu, Guangquan Zhang, Masashi Sugiyama. Butterfly: A Panacea for All Difficulties in Wildly Unsupervised Domain Adaptation, *NeurIPS 2019 Workshop on Learning Transferable Skills*, pp. 1-8, Vancouver, Canada, 8-14 December, 2019. [ERA: A, CORE: A\*]
  7. **Feng Liu**, Guangquan Zhang, Jie Lu. A Novel Fuzzy Neural Network for Unsupervised Domain Adaptation in Heterogeneous Scenarios, *Proceedings of the 2019 IEEE International Conference on Fuzzy Systems* (FUZZ-IEEE 2019), pp. 1-6, New Orleans, USA, 2019. [ERA&CORE: A][**Best Student Paper Award**]
  8. Yiyang Zhang\*, **Feng Liu\***, Zhen Fang, Bo Yuan, Guangquan Zhang, Jie Lu, Clarinet: A One-step Approach Towards Budget-friendly Unsupervised Domain Adaptation, *Proceedings of the 29th International Joint Conference on Artificial Intelligence* (IJCAI 2020), 2020. (\*Equal Contribution). [ERA: A, CORE: A\*]
  9. **Feng Liu**, Guangquan Zhang, Jie Lu, A Novel Non-parametric Two-Sample Test on Imprecise Observations, *Proceedings of the 2020 IEEE International Conference on Fuzzy Systems* (FUZZ-IEEE 2020), online, 19-24 July, 2020. [ERA&CORE: A]

- 
10. **Feng Liu**, Guangquan Zhang, Jie Lu. Unconstrained fuzzy feature fusion for heterogeneous unsupervised domain adaptation, *Proceedings of the 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2018)*, pp. 1-8, Rio de Janeiro, Brazil, 8-13 July, 2018. [ERA&CORE: A]
  11. **Feng Liu**, Guangquan Zhang, Jie Lu. Heterogeneous unsupervised domain adaptation based on fuzzy feature fusion, *Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2017)*, pp. 1-6, Naples, Italy, 9-12 July, 2017. [ERA&CORE: A]
  12. **Feng Liu**, Guangquan Zhang, Anjin Liu, Jie Lu, Discrepancy of Diverse Subsets for Distribution Comparison, *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE-TPAMI)*. [ERA&CORE: A\*, JCR Q1] (under review)
  13. **Feng Liu**, Zhen Fang, Guangquan Zhang, Jie Lu, Take A Closer Look at Kernel Nonparametric Two-sample Tests via the Lebesgue-Besicovitch Differentiation Theorem, *IEEE Transactions on Neural Networks and Learning Systems (IEEE-TNNLS)*. [ERA&CORE: A\*, JCR Q1] (submitted)
  14. **Feng Liu**, Guangquan Zhang, Jie Lu, A Knowledge-Ensemble Model for Heterogeneous Unsupervised Domain Adaptation, *IEEE Transactions on Knowledge and Data Engineering (IEEE-TKDE)*. [ERA: A, CORE: A\*, JCR Q1] (submitted)
  15. Yiyang Zhang\*, **Feng Liu\***, Zhen Fang, Bo Yuan, Guangquan Zhang, Jie Lu, Learning from a Complementary-label Source Domain: Theory and Algorithms, *IEEE Transactions on Neural Networks and Learning Systems*

- 
- (IEEE-TNNLS). (\*Equal Contribution). [ERA&CORE: A\*, JCR Q1] (under review)
16. Hua Zuo, Jie Lu, Guangquan Zhang, **Feng Liu**, Fuzzy Transfer Learning Using an Infinite Gaussian Mixture Model and Active Learning, *IEEE Transactions on Fuzzy Systems* (IEEE-TFS), Vol. 27, no. 2, pp. 291 - 303, 2019. [ERA&CORE: A\*, JCR Q1].
  17. Zhen Fang, Jie Lu, **Feng Liu**, Junyu Xuan, Guangquan Zhang, Open Set Domain Adaptation: Theoretical Bound and Algorithm, *IEEE Transactions on Neural Networks and Learning Systems* (IEEE-TNNLS), 2020. DOI: 10.1109/TNNLS.2020.3017213. [ERA&CORE: A\*, JCR Q1]
  18. Zhen Fang, Jie Lu, **Feng Liu**, Guangquan Zhang. Unsupervised Domain Adaptation with Sphere Retracting Transformation, *Proceedings of the 2019 IEEE International Joint Conference on Neural Networks (IJCNN 2019)*, pp. 1-8, Budapest, Hungary, 14-19 July, 2019. [ERA&CORE: A]
  19. Anjin Liu, Jie Lu, **Feng Liu**, Guangquan Zhang, Accumulating regional density dissimilarity for concept drift detection in data streams, *Pattern Recognition* (PR), Vol. 76, pp. 256-272, 2018. [ERA&CORE: A\*, JCR Q1]
  20. Yi Zhang, Jie Lu, **Feng Liu**, et. al., Does deep learning help topic extraction? A kernel k-means clustering method with word embedding, *Journal of Informetrics*, Vol. 12, no. 4, pp. 1099-1117, 2018. [ERA&CORE: A, JCR Q1]
  21. Li Zhong\*, Zhen Fang\*, **Feng Liu**, Yuan Bo, Guangquan Zhang, Jie Lu, Open Set Domain Adaptation: Theoretical Bound and Algorithm, *IEEE*

---

*Transactions on Neural Networks and Learning Systems* (IEEE-TNNLS).  
(\*Equal Contribution). [ERA&CORE: A\*, JCR Q1] (under review)

22. Qian Zhang, Dianshuang Wu, Jie Lu, **Feng Liu**, Guangquan Zhang, A cross-domain recommender system with consistent information transfer, *Decision Support Systems* (DSS), Vol. 104, pp. 49-63, 2017. [ERA: A\*, JCR Q1]
23. Fan Dong, Jie Lu, Yiliao Song, **Feng Liu**, Guangquan Zhang, A Concept Drift Region-based Data Sample Editing Methodology, *IEEE Transactions on Cybernetics* (IEEE-TCYB). [ERA&CORE: A, JCR Q1] (under review)

## TABLE OF CONTENTS

<b>List of Publications</b>	<b>viii</b>
<b>List of Figures</b>	<b>xxi</b>
<b>List of Tables</b>	<b>xxviii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Motivation . . . . .	3
1.3 Research Questions and Objectives . . . . .	4
1.4 Research Innovation and Contributions . . . . .	8
1.4.1 Research Innovation . . . . .	9
1.4.2 Research Contributions . . . . .	10
1.5 Research Significance . . . . .	12
1.6 Thesis Structure . . . . .	14
<b>2 Literature Review</b>	<b>18</b>
2.1 Transfer Learning . . . . .	18
2.2 Two-sample Test . . . . .	21
2.2.1 F-divergence based Non-parametric Two-sample Tests . . .	22

2.2.2	Function based Non-parametric Two-sample Tests . . . . .	23
2.2.3	Subset based Non-parametric Two-sample Tests . . . . .	25
2.3	Domain Adaptation . . . . .	27
2.3.1	Homogeneous Unsupervised Domain Adaptation . . . . .	28
2.3.2	Heterogeneous Domain Adaptation . . . . .	31
<b>3</b>	<b>Discrepancy of Diverse Subsets: A Non-parametric Two-Sample</b>	
	<b>Test for Low-Dimension Data</b>	<b>37</b>
3.1	Introduction . . . . .	37
3.2	Problem Setting . . . . .	41
3.3	Discrepancy of Diverse Subsets . . . . .	41
3.3.1	Population Form of DDS . . . . .	41
3.3.2	Empirical DDS . . . . .	45
3.3.3	Automatic Selection of Beta . . . . .	47
3.3.4	Consistency of the empirical DDS . . . . .	48
3.3.5	DDS as a Semimetric in $P(X)$ . . . . .	50
3.4	DDS for Two-sample Test . . . . .	51
3.4.1	Asymptotic Distribution of DDS . . . . .	51
3.4.2	Two Hypothesis Tests based on DDS . . . . .	54
3.5	DDS for Unsupervised Domain Adaptation . . . . .	55
3.5.1	Generalization Bound for DDS based Unsupervised Do- main Adaption . . . . .	56
3.5.2	DDS based Unsupervised Domain Adaption . . . . .	60
3.6	Experiments . . . . .	62
3.6.1	Evaluation of DDS for Two-sample Test . . . . .	62



3.6.2	Evaluation of DDS for Unsupervised Domain Adaptation .	69
3.7	Summary . . . . .	74
<b>4</b>	<b>Maximum Mean Discrepancy with Learned Deep Kernels: A</b>	
	<b>Non-parametric Two-Sample Test for High-Dimension Data</b>	<b>76</b>
4.1	Introduction . . . . .	76
4.2	Concepts and Notations . . . . .	79
4.3	Limits of Simple Kernels . . . . .	84
4.4	Relationship to Classifier-Based Tests . . . . .	85
4.5	Learning Deep Kernels . . . . .	88
4.6	Theoretical Analysis . . . . .	89
4.7	Experimental Results . . . . .	92
4.7.1	Comparison on Benchmark Datasets . . . . .	92
4.7.2	Ablation Study . . . . .	101
4.7.3	Interpretability on <i>CIFAR-10</i> vs <i>CIFAR-10.1</i> . . . . .	102
4.8	Summary . . . . .	104
<b>5</b>	<b>Butterfly: A One-step Approach towards Wildly Unsupervised</b>	
	<b>Domain Adaptation</b>	<b>107</b>
5.1	Introduction . . . . .	107
5.2	Wildly Unsupervised Domain Adaptation . . . . .	111
5.2.1	Nature of WUDA . . . . .	112
5.2.2	WUDA ruins HoUDA Methods . . . . .	113
5.3	Butterfly: Towards Robust One-step Approach . . . . .	114
5.3.1	Loss Function in Butterfly . . . . .	115
5.3.2	Training Procedure of Butterfly . . . . .	115

5.4	Butterfly vs. Two-step Approach . . . . .	118
5.4.1	Two-step Approach (A Compromise Solution) . . . . .	119
5.4.2	One-step Approach (Butterfly) . . . . .	119
5.5	Theoretical Analysis . . . . .	120
5.5.1	WUDA ruins HoUDA Methods . . . . .	122
5.5.2	Two-step Approach is a Compromise Solution . . . . .	123
5.5.3	Why does Butterfly Eliminate Noise Effect? . . . . .	124
5.5.4	Principle-guided Butterfly . . . . .	129
5.5.5	A Generalization Bound for WUDA . . . . .	130
5.6	Comparison to Related Works . . . . .	133
5.7	Experiments . . . . .	134
5.7.1	Simulated WUDA Tasks . . . . .	134
5.7.2	Real-world WUDA Tasks . . . . .	136
5.7.3	Baselines . . . . .	137
5.7.4	Network Structure and Optimizer . . . . .	138
5.7.5	Experimental Setup . . . . .	139
5.7.6	Results on Simulated WUDA Tasks . . . . .	140
5.7.7	Results on Real-world WUDA Tasks . . . . .	144
5.7.8	Ablation Study . . . . .	144
5.8	Summary . . . . .	148
<b>6</b>	<b>Heterogeneous Domain Adaptation: An Unsupervised Approach and Its Theoretical Guarantees</b>	<b>149</b>
6.1	Introduction . . . . .	149
6.2	Problem Setting and Notations . . . . .	154

6.3	Heterogeneous Unsupervised domain adaptation . . . . .	155
6.3.1	Unsupervised Knowledge Transfer Theorem for HeUDA .	156
6.3.2	Principal Angle-based Measurement between Heterogeneous Feature Spaces . . . . .	160
6.3.3	GLG: The Proposed HeUDA Method . . . . .	163
6.3.4	Discussion of Definitions and Theorems . . . . .	169
6.3.5	Limitation of GLG . . . . .	171
6.4	Optimization of GLG . . . . .	171
6.4.1	Microscopic Analysis of an Eigen Dynamic System . . . . .	172
6.4.2	Gradients of Cost Function I . . . . .	173
6.4.3	A Hybrid Optimization Method for GLG . . . . .	175
6.5	Experiments . . . . .	176
6.5.1	Datasets for HeUDA . . . . .	176
6.5.2	Experimental Setup . . . . .	180
6.5.3	Experiment I: RMG . . . . .	184
6.5.4	Experiment II: Overall comparisons . . . . .	187
6.6	Summary . . . . .	191
<b>7</b>	<b>Shared Fuzzy Equivalence Relations: A Heterogeneous Unsupervised Domain Adaptation Approach for Imbalanced Domains</b>	<b>192</b>
7.1	Introduction . . . . .	192
7.2	Concepts and Notations . . . . .	195
7.2.1	Fuzzy Geometry . . . . .	195
7.2.2	Fuzzy Equivalence Relation . . . . .	197
7.3	Similarity between Fuzzy Vectors . . . . .	200

7.3.1	A Metric on Fuzzy Geometry . . . . .	200
7.3.2	Similarity between Fuzzy Vectors . . . . .	204
7.4	F-HeUDA via Shared Fuzzy Equivalence Relations . . . . .	205
7.4.1	Theoretical Guarantees . . . . .	206
7.4.2	Shared Fuzzy Equivalence Relations (SFER) . . . . .	211
7.4.3	Learning Process of SFER . . . . .	212
7.4.4	F-HeUDA via SFER (F-HeUDA) . . . . .	214
7.5	Experiments . . . . .	216
7.5.1	Dataset Description and Parameters Setting . . . . .	216
7.5.2	Prediction Performance . . . . .	218
7.5.3	Convergence of Learning Algorithm . . . . .	221
7.5.4	User-oriented Decision Making Pattern . . . . .	222
7.6	Summary . . . . .	224
<b>8</b>	<b>Fuzzy-relation Neural Networks: A Multi-source Heterogeneous</b>	
	<b>Unsupervised Domain Adaptation Approach</b>	<b>226</b>
8.1	Introduction . . . . .	226
8.2	Concepts and Problem Setting . . . . .	230
8.2.1	Similarity between Fuzzy Vectors . . . . .	230
8.2.2	Fuzzy Equivalence Relations and Partitioning of Fuzzy Sets	232
8.2.3	Multi-source Unsupervised Domain Adaptation . . . . .	233
8.3	Shared Fuzzy Equivalence Relations for Multi-source Domains .	234
8.3.1	Theoretical Guarantees for Multi-source SFER . . . . .	236
8.3.2	Multi-source SFER . . . . .	239
8.3.3	Learning Process of Multi-source SFER . . . . .	239

8.4	Shared Fuzzy Equivalence Relations Neural Network for Multi-source HeUDA . . . . .	243
8.4.1	Structure of the Proposed Neural Network . . . . .	243
8.4.2	Loss Function of the Proposed Neural Network . . . . .	247
8.4.3	Learning Process of the Proposed Neural Network . . . . .	248
8.5	Experiments . . . . .	250
8.5.1	Datasets and Tasks . . . . .	250
8.5.2	Experimental Setup . . . . .	251
8.5.3	Experiment I: Classification Accuracy . . . . .	253
8.5.4	Experiment II: Stability Analysis . . . . .	255
8.5.5	Experiment III: Parameters Sensitivity . . . . .	258
8.6	Summary . . . . .	260
<b>9</b>	<b>Conclusion and Future Study</b>	<b>263</b>
9.1	Conclusions . . . . .	263
9.2	Future Study . . . . .	267
<b>A</b>	<b>Appendix</b>	<b>270</b>
A.1	Appendix of Chapter 3 . . . . .	270
A.1.1	A Corollary of Radon-Nikodym Theorem . . . . .	270
A.1.2	Proofs of Theorems . . . . .	271
A.2	Appendix of Chapter 4 . . . . .	277
A.2.1	Preliminaries . . . . .	277
A.2.2	Main Results . . . . .	279
A.2.3	Uniform Convergence Results . . . . .	281
A.2.4	Constructing Appropriate Kernels . . . . .	282

A.2.5	Miscellaneous Results . . . . .	287
A.3	Appendix of Chapter 5 . . . . .	288
A.3.1	Review of Generation of Noisy Labels . . . . .	288
A.3.2	Proofs . . . . .	290
A.4	Appendix of Chapter 6 . . . . .	301
A.4.1	Proof of Theorem 6.1 . . . . .	301
A.4.2	Proof of Lemma 6.1 . . . . .	302
A.4.3	Proof of Theorem 6.2 . . . . .	302
A.4.4	Proof of Theorem 6.3 . . . . .	302
A.4.5	Proof of Theorem 6.4 . . . . .	303
A.4.6	Proof of Lemma 6.2 . . . . .	304
<b>Bibliography</b>		<b>306</b>

## LIST OF FIGURES

FIGURE	Page
1.1 Thesis structure . . . . .	17
3.1 The proposed DDS based unsupervised domain adaptation (DSA). Red circles with the solid line represent the features of both domains, and blue circles with solid line represent in-sample (source domain) outputs of DSA, and blue circles with dash line represent out-sample (target domain) outputs of DSA. Other circles represent hidden neu- rons of DSA. In the third layer of DSA, representations of instances of both domains are obtained: $T(X_s)$ and $T(X_t)$ . . . . .	62
4.1 Blob dataset (a), with contours of Gaussian kernel (b) and deep kernel (c) evaluated at 9 locations (contour values are 0.7, 0.8 and 0.9). Each distribution has 9 modes; the central modes have the same shape, but $\mathbb{Q}$ has a different shape at each other mode. A Gaussian kernel (b) compares points isotropically throughout the space; contours show $k(x, \mu)$ for each mode $\mu$ . A deep kernel (c) learned by our methods compares points differently in different locations, allowing better identification of differences between $\mathbb{P}$ and $\mathbb{Q}$ . . . . .	78

4.2	Results on <i>Blob-S</i> and <i>Blob-D</i> given $\alpha = 0.05$ ; see Section 4.7 for details. $n_b$ is the number of samples at each mode, so $n_b = 100$ means drawing 900 samples from each of $\mathbb{P}$ and $\mathbb{Q}$ . We report, when increasing $n_b$ , (a) average test power, (b) standard deviation of test power, (c) the value of $\hat{J}_\lambda$ , and (d) average type-I error. (a), (b) and (c) are on <i>Blob-D</i> , and (d) is on <i>Blob-S</i> . Shaded regions show standard errors for the mean, and the black line shows $\alpha$ . . . . .	85
4.3	Results on <i>HDGM-S</i> and <i>HDGM-D</i> for $\alpha = 0.05$ (black line). Left: average test power (a) and Type I error (b) when increasing the number of samples $N$ , keeping $d = 10$ . Right: average test power (c) and Type I error (d) when increasing the dimension $d$ , keeping $N = 4000$ . Shaded regions show standard errors for the mean. . . . .	92
4.4	The structure of $\phi_\omega$ in MMD-D on <i>MNIST</i> . The kernel size of each convolutional layer is 3; stride (S) is set to 2; padding (P) is set to 1. We do not use dropout. Best viewed zoomed in. . . . .	95
4.5	The structure of classifier $F$ in C2ST-S and C2ST-L on <i>MNIST</i> . The kernel size of each convolutional layer is 3; stride (S) is set to 2; padding (P) is set to 1. We do not use dropout. In the first layer, we will convert the <i>CIFAR</i> images from $32 \times 32 \times 3$ to $64 \times 64 \times 3$ . Best viewed zoomed in. . . . .	95
4.6	The structure of $\phi_\omega$ in MMD-D on <i>CIFAR</i> . The kernel size of each convolutional layer is 3; stride (S) is set to 2; padding (P) is set to 1. We do not use dropout in all layers. In the first layer, we will convert the <i>CIFAR</i> images from $32 \times 32 \times 3$ to $64 \times 64 \times 3$ . Best viewed zoomed in.	95



4.7	The structure of classifier $F$ in C2ST-S and C2ST-L on <i>CIFAR</i> . The kernel size of each convolutional layer is 3; stride (S) is set to 2; padding (P) is set to 1. We do not use dropout. Best viewed zoomed in.	95
4.8	The best test locations (learned by an ME test with $L = 1$ ) from 10 experiments on <i>CIFAR-10</i> vs <i>CIFAR-10.1</i> . Average rejection rate is 0.415. . . . .	104
4.9	The best test locations (learned by an ME test, $L = 1$ , with a deep kernel optimized for an MMD test) from 10 experiments on <i>CIFAR-10</i> vs <i>CIFAR-10.1</i> . Average rejection rate is 0.637. . . . .	105
4.10	The best test locations (selected among existing images with our learned deep kernel, $L = 1$ ) from 10 experiments on <i>CIFAR-10</i> vs <i>CIFAR-10.1</i> . Average rejection rate is 0.653. . . . .	105
5.1	Wildly unsupervised domain adaptation (WUDA). The blue line denotes that HoUDA transfers knowledge from clean source data ( $P_s$ ) to unlabeled target data ( $P_{x_t}$ ). However, perfectly clean data is hard to acquire. This brings <i>wildly unsupervised domain adaptation</i> (WUDA), namely transferring knowledge from noisy source data ( $\tilde{P}_s$ ) to unlabeled target data ( $P_{x_t}$ ). Note that label corruption process (black dash line) is unknown in practice. To handle WUDA, a compromise solution is a two-step approach (green line), which sequentially combines label-noise algorithms ( $\tilde{P}_s \rightarrow \hat{P}_s$ , label correction) and existing HoUDA ( $\hat{P}_s \rightarrow P_{x_t}$ ). This chapter proposes a robust one-step approach called Butterfly (red line, $\tilde{P}_s \rightarrow P_{x_t}$ directly), which eliminates noise effects from $\tilde{P}_s$ . . . . .	108

- 5.2 WUDA ruins representative HoUDA methods. Representative HoUDA methods includes *deep adaptation network* (DAN, a IPM based method [Long et al., 2015]), *domain-adversarial neural network* (DANN, a adversarial training based method [Ganin et al., 2016a]), *asymmetric tri-training domain adaptation* (ATDA, a pseudo-label based method [Saito et al., 2017]) and *transferable curriculum learning* (TCL, a robust HoUDA method [Shu et al., 2019]). B-Net is our proposed WUDA method. We report target-domain accuracy of all methods when the noise rate of source domain changes (a) from 5% to 70% (symmetry-flip noise) and (b) from 5% to 45% (pair-flip noise). Clearly, when the noise rate of source domain increases, target-domain accuracy of representative HoUDA methods drops quickly while that of B-Net keeps stable consistently. . . . . 109
- 5.3 Butterfly Framework. Two networks ( $F_1$  and  $F_2$ ) in Branch-I are jointly trained on noisy source data and pseudo-labeled target data (mixture domain). Two networks in Branch-II ( $F_{t1}$  and  $F_{t2}$ ) are trained on pseudo-labeled target data. By using dual-checking principle, Butterfly checks high-correctness data out from both mixture and pseudo-labeled target data. After cross-propagating checked data, Butterfly can obtain high-quality *domain-invariant representations* (DIR) and *target-specific representations* (TSR) simultaneously in an iterative manner. Note that the interaction between DIR and TSR happens via the shared CNN. Besides, in the first training epoch, since we do not have any pseudo-labeled target data, we need to use noisy source data as the pseudo-labeled target data, which follows [Saito et al., 2017]. . 112

5.4	Visualization of <i>MNIST</i> and <i>SYND</i> . . . . .	136
5.5	Visualization of <i>Bing</i> , <i>Caltech256</i> , <i>ImageNet</i> and <i>SUN</i> (taking “horse” as the common class). . . . .	137
5.6	The architecture of B-Net for digit WUDA tasks $SYND \leftrightarrow MNIST$ . We added BN layer in the last convolution layer in CNN and FC layers in $F_1$ and $F_2$ . We also used dropout in the last convolution layer in CNN and FC layers in $F_1$ , $F_2$ , $F_{t1}$ and $F_{t2}$ (dropout probability is set to 0.5). . . . .	138
5.7	The architecture of B-Net for (a) human-sentiment WUDA tasks and (b) real-world WUDA tasks. We added BN layer in the first FC layers in $F_1$ and $F_2$ . We also used dropout in the first FC layers in $F_1$ , $F_2$ , $F_{t1}$ and $F_{t2}$ (dropout probability is set to 0.5). . . . .	139
5.8	Target-domain accuracy vs. number of epochs on four $SYND \rightarrow MNIST$ WUDA tasks. . . . .	141
5.9	Target-domain accuracy vs. number of epochs on four $MNIST \rightarrow SYND$ WUDA tasks. . . . .	142
6.1	The progress of the GLG method. The original source and target domains come from the same underlying domain (e.g., classifying Latin sentences or analyzing human sentiment). However, the underlying domain is hard to observe and we can only observe its projection/representation on two (or more) domains, e.g., two heterogeneous domains in this figure. . . . .	153
7.1	Relationships among $\sup\{D_\lambda(u, v) : D_\lambda(u, v) \in \Omega(\lambda)\}$ , $d(u, \bar{A}_j(\lambda))$ , $d(v, \bar{A}_i(\lambda))$ and $d(A_i, A_j)$ . . . . .	202

7.2	Traditional fuzzy equivalence relations v.s. SFER. In subfigure (a), two domains clearly cannot use the same $\alpha$ to obtain the same number of clusters. But, in SFER, two domains have a much bigger probability of using the same $\alpha$ to obtain the same number of clusters. . . . .	212
7.3	The performance of each method (mean accuracy and standard deviation) on 4 tasks. In each subfigure, the minimum mean accuracy is set as 0.4. . . . .	220
7.4	The convergence of Algorithm 7.1 on four tasks. Subfigures (a)-(d) illustrate the value of the cost function $J_1$ and subfigures (e)-(f) illustrate the $r_i$ of $R_{TD}^M$ of the source domain and the target domain. . . .	222
8.1	Three scenarios for the <i>unsupervised domain adaptation</i> (UDA) problem. Rectangles represent the labeled source data and triangle represents the unlabeled target data. Labeled source data may come from a) a single source domain (i.e., single-source scenario) or b) multiple source domains whose feature spaces have the same dimension (i.e., multi-homogeneous-source scenario) or c) multiple source domains whose feature spaces have different dimensions (i.e., multi-heterogeneous-source scenario). Given a target domain, since we probably have many different-dimension source domains (i.e., multi-heterogeneous-source scenario), the third scenario is more general than the other scenarios. . . . .	227

8.2	Target domain has 7 features while <i>source domain 1</i> has 6 and <i>source domain 2</i> has 5. Given the same $\alpha$ , traditional fuzzy equivalence relations can only simultaneously cluster 1) features in <i>source domain 1</i> as 3 categories, and 2) features in <i>source domain 2</i> as 2 categories, and 3) features in target domain as 5 categories. However, multi-source shared fuzzy equivalence relations can simultaneously cluster features in <i>source domain 1</i> , <i>source domain 2</i> and target domain as 2 categories. . . . .	235
8.3	Network struture of SFERNN. SFERNN is a five-layer neural network containing $c$ source branches and one target branch ( $c = 2$ in this figure). Network structure (i.e., $N_l$ in this figure and how to connect two adjacent layers) of SFERNN is confirmed by MsSFER. Loss function of SFERNN is composed of two parts. The first part represents cross-entropy loss on labeled data from $c$ source domains. The second part represents distributional discrepancy (MMD) between source domains and target domain. . . . .	244
8.4	Analysis of parameters' sensitivity on 3 tasks. Learning rate represents $\eta$ in Algorithm 8.2 and Lambda represents $\lambda_2$ in loss function of SFERNN. . . . .	262

## LIST OF TABLES

<b>TABLE</b>	<b>Page</b>
<p>3.1 Synthetic data set results (%). The null hypothesis <math>H_0</math> is <math>p = q</math>. The <math>H_1^{\Delta_i}</math> indicates the percentage of the test reject the null hypothesis, while the magnitude between <math>p</math> and <math>q</math> is <math>\Delta_i</math>. Since <math>\Delta_1</math> is equal to 0, <math>H_1^{\Delta_1}</math> is the Type I error (the lower the better), while <math>H_1^{\Delta_2}</math> and <math>H_1^{\Delta_3}</math> are the Type II error (the lower the better). Bold values represent the lowest error. . . . .</p>	67
<p>3.2 Average Type I error (%) and corresponding rankings. This table shows the average Type I error for each test and the average values of Type I errors obtained under <math>m_1, m_2, m_3</math> (see Table 3.1). Bold values represent the lowest error. . . . .</p>	68
<p>3.3 Average Type II error (%) and corresponding rankings. This table shows the average Type II error for each test and the average values of Type II errors obtained under <math>m_1, m_2, m_3</math> (see Table 3.1). Bold values represent the lowest error. . . . .</p>	68

3.4	Higgs data set results (%) and corresponding rankings. Real-I has both two-sample sets drawn from $P$ distribution. Real-II has both two sample sets drawn from $Q$ distribution. Real-III has one sample set drawn from $P$ and the other drawn from $Q$ . Bold values represent the lowest error and lowest running time. . . . .	69
3.5	Real-world datasets for transfer learning . . . . .	70
3.6	Classification accuracy for DDS-based domain adaptation and MMD-based domain adaptation on 28 transfer recognition tasks (object recognition (the first 8 rows) and face recognition (the other rows)). The results show that DDS outperformed MMD in 26 transfer tasks. Bold values represent the highest accuracy. Please note that 1-NN means the 1 nearest neighbor and NN means the neural network. . .	75
4.1	Specifications of $\mathbb{P}$ and $\mathbb{Q}$ of synthetic datasets. $\mu_1^b = [0, 0], \mu_2^b = [0, 1], \mu_3^b = [0, 2], \dots, \mu_8^b = [2, 1], \mu_9^b = [2, 2]$ (same with Figure 4.1a). $\mu_1^h = \mathbf{0}_d, \mu_2^h = 0.5 \times \mathbf{1}_d, I_d$ is an identity matrix with size $d$ . $\Delta_i^b = -0.02 - 0.002 \times (i - 1)$ if $i < 5$ and $\Delta_i^b = 0.02 + 0.002 \times (i - 6)$ if $i > 5$ . if $i = 5, \Delta_i^b = 0$ (same with Figure 4.1a). $\Delta_1^h$ and $\Delta_2^h$ are set to 0.5 and $-0.5$ , respectively. . . . .	97
4.2	Higgs ( $\alpha = 0.05$ ): average test power $\pm$ standard error for $N$ samples. Bold represents the highest mean per row. . . . .	97

4.3	Results on <i>Higgs</i> ( $\alpha = 0.05$ ). We report average Type I error on <i>Higgs</i> dataset when increasing number of samples ( $N$ ). Note that, in <i>Higgs</i> , we have two types of Type I errors: 1) Type I error when two samples drawn from $\mathbb{P}$ (no Higgs bosons) and 2) Type I error when two samples drawn from $\mathbb{Q}$ (having Higgs bosons). Type I reported here is the average value of 1) and 2). Since Type I error reported here is the average value of two average Type I errors, we do not report standard errors of the average Type I error in this table. . . . .	98
4.4	<i>MNIST</i> ( $\alpha = 0.05$ ): average test power $\pm$ standard error for comparing $N$ real images to $N$ DCGAN samples. . . . .	98
4.5	Results on <i>MNIST</i> given $\alpha = 0.05$ . We report average Type I error $\pm$ standard errors on real- <i>MNIST</i> vs. real- <i>MNIST</i> when increasing number of samples ( $N$ ). . . . .	98
4.6	<i>CIFAR-10.1</i> ( $\alpha = 0.05$ ): mean rejection rates. . . . .	101
4.7	Mean test power on <i>Blob</i> ( $n_b = 40$ ), <i>HDGM</i> ( $N = 4000, d = 10$ ), <i>Higgs</i> ( $N = 3000$ ) and <i>MNIST</i> ( $N = 400$ ) for $\alpha = 0.05$ . See Section 4.7.2 for the naming scheme; S+C corresponds to C2ST-S, L+C to C2ST-L, and D+J to MMD-D. L+M is the method proposed by <a href="#">Kirchler et al. [2019]</a> . . . . .	103
4.8	Paired t-test results ( $\alpha = 0.05$ ) for the results of Section 4.7.1. For <i>HDGM</i> , we fix $d = 10$ (corresponding to Figure 4.3a). $\checkmark$ indicates MMD-D achieved statistically significantly higher mean test power than the other method, $\times$ that it did not. . . . .	103
5.1	Target-domain accuracy on 8 digit WUDA tasks ( <i>SYND</i> $\leftrightarrow$ <i>MNIST</i> ). Bold value represents the highest accuracy in each row. . . . .	140



5.2	Target-domain accuracy on 12 human-sentiment WUDA tasks with the 20% noise rate. Bold values mean the highest values in each row.	143
5.3	Target-domain accuracy on 12 human-sentiment WUDA tasks with the 45% noise rate. Bold values mean the highest values in each row.	143
5.4	Target-domain accuracy on 3 real-world WUDA tasks. The source domain is the <i>Bing</i> dataset that contains noisy information from the Internet. Bold value represents the highest accuracy in each row.	144
5.5	Results of ablation study. Average target-domain accuracy on 8 simulated digit WUDA tasks ( <i>Digit</i> ), 24 simulated human-sentiment WUDA tasks ( <i>Sentiment</i> ) and 3 real-world WUDA tasks ( <i>Real-world</i> ). Bold value represents the highest accuracy in each row.	147
6.1	Description of the original datasets.	178
6.2	Transfer tasks (10 tasks in total).	179
6.3	Same-domain accuracy of each target domain using 5-fold SVM.	183
6.4	The classification results for RMG and CM.	185
6.5	The results of the MMD test for the mapped and adapted domains in two extreme situations (lowest and highest accuracy) of task CD2CO among 50-time experiments.	186
6.6	The classification results (AVG $\pm$ STD) for GLG and benchmark models. Bold values represent the lowest average accuracy in each task.	190
7.1	The overall performance of each method on four tasks.	219
7.2	The overall STD value of each method on four tasks.	221
7.3	The prediction performance of F-HeUDA When changing $\alpha$ .	223

8.1	Description of the three datasets. . . . .	251
8.2	Classification accuracy of the baselines and SFERNN on the Australian credit task (Task 1). SFERNN can extract useful information from both the <i>source domains</i> (SDs) and obtain better average accuracy on the <i>target domain</i> (TD) than all the baselines, no matter which source domain these baselines select. . . . .	254
8.3	Classification accuracy of the baselines and SFERNN on the German credit task (Task 2). SFERNN can extract useful information from both the <i>source domains</i> (SDs) and obtain a better average accuracy on the <i>target domain</i> (TD) than all the baselines, no matter which source domain these baselines select. . . . .	256
8.4	Classification accuracy of the baselines and SFERNN on the Japanese credit task (Task 3). SFERNN can extract useful information from both the <i>source domains</i> (SDs) and obtain a better average accuracy on the <i>target domain</i> (TD) than all the baselines, no matter which source domain these baselines select. . . . .	257
8.5	STD of classification accuracy of baselines and SFERNN on the Australian credit task (Task 1). SFERNN can use knowledge in the <i>source domains</i> (SDs) to obtain more stable accuracy across 50 experiments on the <i>target domain</i> (TD) than most baselines. . . . .	259
8.6	The STD of the classification accuracy of the baselines and SFERNN on the German credit task (Task 2). SFERNN can use knowledge in <i>source domains</i> (SDs) to obtain more stable accuracy across 50 experiments on the target domain (TD) than most baselines. . . . .	260

8.7	The STD of classification accuracy of the baselines and SFERNN on the Japanese credit task (Task 3). SFERNN can use knowledge in <i>source domains</i> (SDs) to obtain more stable accuracy across 50 experiments on the <i>target domain</i> (TD) than most baselines. . . . .	261
-----	---	-----